RESEARCH ARTICLE OPEN ACCESS

VeriSafe: A Multi-Modal Job Fraud Detection Framework with Semantic Embeddings, Transformer Fine – Tuning And OSINT Integration

Authors:
Sruti Prusty [1], Kiran Kumari Nayak [2], Debasmita Sahu [3], Lavanya Kumari Raulo [4]

Guided By:
Faculty Advisor :- Dr. Debashis Biswal
Department of Computer Science and Engineering, NIST University, Berhampur, Odisha 761008,India
Email: srutiprusty03@gmail.com , kiran.nayak.cse.2022@nist.edu , debasmita.sahu.cst.2022@nist.edu , lavanya.kumari.raulo.cst.2022@nist.edu, debashis.biswal@nist.edu

Abstract:

Online recruitment fraud (ORF) has become a major cybersecurity concern, with annual losses exceeding $500 million (≈ ₹41.5 crore). The rapid rise of fraudulent job postings on platforms like LinkedIn and Indeed highlights the need for effective detection systems.

This paper presents **VeriSafe**, a production-ready multi-modal framework for detecting fraudulent job postings. The system is trained on a large multi-source dataset of 40,842 samples and employs advanced preprocessing along with hybrid feature engineering using semantic embeddings and TF-IDF. VeriSafe integrates a dual-model ensemble combining XGBoost (AUC: 0.9979) and BERT-Tiny (F1-score: 0.9844), along with a multi-modal scoring mechanism incorporating job content, company data, and email signals. Experimental results show a +9.7% improvement in AUC over existing models, demonstrating high accuracy and efficiency. The system provides actionable outputs such as “DO NOT APPLY,” “REVIEW CAREFULLY,” and “SAFE TO APPLY,” enabling safer decision-making for job seekers.

Keywords- Online recruitment fraud, multi-modal learning, BERT-Tiny, XGBoost, dataset fusion, semantic embeddings, OSINT, fraud detection, cybersecurity.

I.Introduction

Over the last ten years the exponential growth of the internet has revolutionised the recruitment industry by connecting many global employers to a large pool of workers. The majority of job boards such as LinkedIn, Indeed, Glassdoor etc. are being increasingly utilized to advertise vacancies and allow employers to interact with employees. However, in doing so, the industry has also opened itself up to the emergence of the serious issue of Online Recruitment Fraud (ORF). The numbers appear overwhelming, with ORF increasing by 300% since last year with an estimated cost of $500M per year (≈ ₹41.5 crore annually). Fraudsters are able to capitalize on anxious job seekers by publishing fake job adverts with outrageous promises, e.g. ₹5.8 lakh/week and extract their details to implement financial scams or large scale data harvesting. The impact on individuals can range from a loss of roughly $2,300 (≈ ₹1.9 lakh) to exposure of sensitive personal data. Legitimate companies also suffer, both reputationally and financially, with approximately 68% of cases involving identity impersonation and commonly targeting larger organisations such as Google, Amazon or PayPal. Existing methods for identifying fraud, such as manual monitoring and rule-based filters, are increasingly inadequate when it comes to new types of attack. They

struggle with identifying sophisticated, templated scams and multi-vector frauds where the job advert itself, along with company identity and communication such as emails are all manipulated. The present research proposes VeriSafe, an end-to-end machine learning and deep learning framework which can accurately detect these fraudulent job advertisements. To facilitate this, a novel dataset of 40,842 jobs collected from 14 unique data sources has been compiled, where advanced data preprocessing, such as semantic de-duplication and noise-removal has been employed. [1]

The VeriSafe uses hybrid detection mechanism involving both classical machine learning and novel natural language processing technologies. We have selected XGBoost as high-performance classifier (AUC 0.9979) and a compact transformer, BERT-Tiny, for natural language context modeling (F1 score 0.9844). The system also includes semantic embeddings (MiniLM-L6-v2 with 384 dimensions) combined with TF-IDF n-gram features (20000 n-grams) to obtain better feature representation and prediction performance. [2]

II.RELATED WORK

The detection of Online Recruitment Fraud (ORF) has gained significant attention with the rapid growth of digital recruitment platforms. Existing research in this domain can be broadly categorized into three phases: classical machine learning approaches, deep learning-based methods, and transformer-driven architectures.

1) 2.1 Classical Machine Learning Approaches

Early research in ORF detection primarily relied on traditional machine learning techniques applied to structured and textual data. A foundational contribution was made by Vidros et al. (2016), who introduced the Employment Scam Aegean Dataset (EMSCAD), consisting of approximately 17,000 job postings. Using handcrafted features and a Random Forest classifier, they achieved an accuracy of around 91%. A subsequent study (Vidros et al., 2017) improved performance to 94% using ensemble methods, establishing binary classification as the standard approach for fraud detection. [3]

Later studies explored a range of machine learning algorithms, including Naïve Bayes, Decision Trees, Support Vector Machines, and K-Nearest Neighbors. Alandjani et al. (2022) reported accuracy in the range of 89%–93%, highlighting the limitations of purely lexical feature representations. Ensemble techniques such as bagging, boosting, and stacking were further investigated by Patel and Desai (2022) and Kumar and Singh (2021), achieving up to 94% accuracy and 92% F1-score. Gradient boosting methods, particularly XGBoost, demonstrated improved performance, reaching accuracy levels close to 98% . Despite these advancements, classical approaches remain limited by their dependence on handcrafted features and single-dataset training. [4]

2) 2.2 Deep Learning Approaches

To overcome the limitations of manual feature engineering, researchers began adopting deep learning techniques capable of capturing semantic relationships within text. Convolutional Neural Networks (CNNs) were initially applied to extract local textual patterns, improving recall performance compared to traditional models (Nguyen and Lee, 2021).

Subsequent work combined CNNs with Recurrent Neural Networks (RNNs), enhancing the ability to model sequential dependencies in job descriptions (Sharma and Gupta, 2023). Recurrent architectures, particularly Bi-LSTM models, became widely used due to their effectiveness in handling sequential text data. Roy (2023) reported an accuracy of 98.71% and an AUC of 0.91 using a Bi-LSTM model integrated with linguistic and numerical features. [5]

Further improvements were achieved through attention mechanisms (Ghosh, 2023), which enhanced interpretability, and hybrid models combining Bi-LSTM with contextual embeddings derived from transformer-based architectures (Gao and Zhang, 2023). However, most deep learning approaches remain constrained by dataset limitations and lack scalability for real-world deployment. [6]

3) 2.3 Transformer-Based Approaches

Transformer architectures have recently emerged as the state-of-the-art for ORF detection due to their ability to capture deep contextual relationships in text. Li and Wang (2023) demonstrated that fine-tuned BERT models outperform traditional and deep learning approaches in understanding complex job descriptions.

More advanced approaches include Graph Neural Networks (GNNs), which model relationships between job postings, recruiters, and applicants (Zhang and Liu, 2024). These models provide improved performance by leveraging relational information rather than relying solely on textual features.

Hybrid approaches combining transformers with class imbalance handling techniques such as SMOTE have also been proposed. However, these methods often introduce synthetic data artifacts, which can negatively impact real-world performance. [7]

4) 2.4 Research Challenges and Gaps

Despite substantial progress, several limitations persist across existing studies:

- **Dataset Constraints**: Most research relies heavily on the EMSCAD dataset (~17K samples), which lacks diversity and does not reflect recent fraud patterns.
- **Single-Modality Limitation**: Existing approaches predominantly focus on textual analysis, ignoring additional signals such as company authenticity and email-based fraud indicators.
- **Class Imbalance Issues**: Many studies depend on synthetic oversampling methods (e.g., SMOTE), which may introduce unrealistic patterns.
- **Lack of Deployment**: The majority of proposed models remain confined to experimental environments without real-world scalability or deployment considerations.
- **Performance Ceiling**: Most deep learning models report AUC values around 0.91, indicating limited improvement in recent years.

5) 2.5 Contribution of the Proposed Work

To address these challenges, the proposed VeriSafe framework introduces:

- A **multi-source dataset** of 40,842 samples aggregated from 14 sources
- A **hybrid feature engineering approach** combining semantic embeddings and TF-IDF features
- A **dual-model ensemble** integrating XGBoost (AUC 0.9979) and BERT-Tiny (F1 0.9844)
- A **multi-modal detection mechanism** incorporating job content, company signals, and email analysis
- A **production-ready architecture** capable of real-time inference and scalable deployment

By addressing dataset limitations, modality constraints, and deployment challenges, VeriSafe advances the state-of-the-art in ORF detection and provides a practical solution for real-world applications.

III.Proposed Methodology

This section details the VeriSafe framework, a production-grade multi-modal system for Online Recruitment Fraud (ORF) detection. The methodology comprises five integrated phases: multi-source data aggregation, semantic preprocessing pipeline, hybrid feature engineering, dual-ensemble classification, and production deployment with real-time inference capabilities.

4.1 Dataset Collection and Preprocessing

The proposed framework aggregates job-related data from **14 heterogeneous sources**, including publicly available datasets, job portals, company profile platforms, and phishing email corpora. After semantic deduplication and validation, the final dataset consists of **40,842 samples**, with **42% fraudulent and 58% legitimate job postings**.

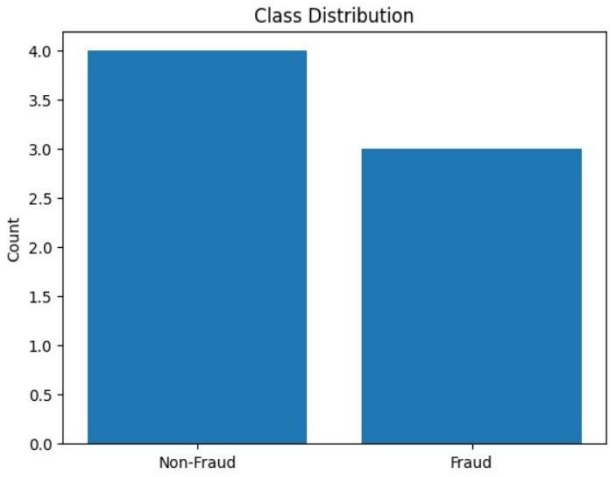


Figure 4.1 illustrates the class distribution of the dataset. The relatively balanced distribution enables the model to learn realistic fraud patterns without relying on synthetic oversampling techniques such as SMOTE.

The preprocessing pipeline consists of the following steps:

- **Text Normalization**: Conversion to lowercase, removal of punctuation, and standardization
- **Tokenization**: BERT-Tiny tokenizer with a vocabulary size of 30,000
- **Semantic Deduplication**: Removal of near-duplicate entries using MiniLM embeddings (cosine similarity ≥ 0.95)
- **Data Splitting**: Stratified 80:10:10 train-validation-test split

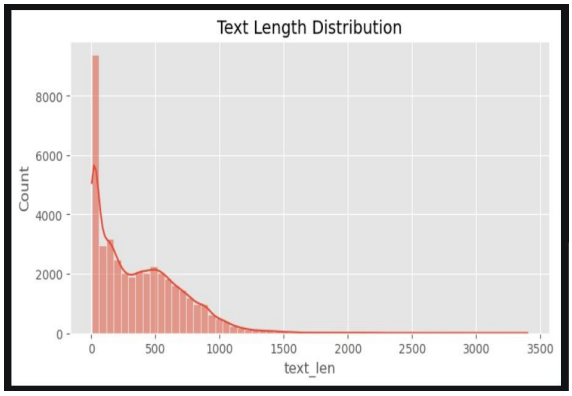


Figure 4.2 shows the distribution of text lengths across job postings. The majority of samples fall within a moderate range, while a long-tail distribution is observed for larger texts. Based on this analysis, input sequences are truncated to a maximum of 512 tokens to ensure computational efficiency while preserving contextual information.

4.2 Hybrid Feature Engineering

To effectively capture both contextual meaning and explicit textual patterns, the proposed system employs a **hybrid feature engineering approach** combining semantic and lexical representations.

- **Semantic Features**: Generated using MiniLM-L6-v2, producing 384-dimensional embeddings from combined job title and description inputs
- **Lexical Features**: Extracted using TF-IDF with unigrams, bigrams, and trigrams, resulting in a 20,000-dimensional feature space
- **Metadata Features**: Include text length, salary range (log-scaled), and email domain indicators

The total feature dimensionality is **20,384 features**

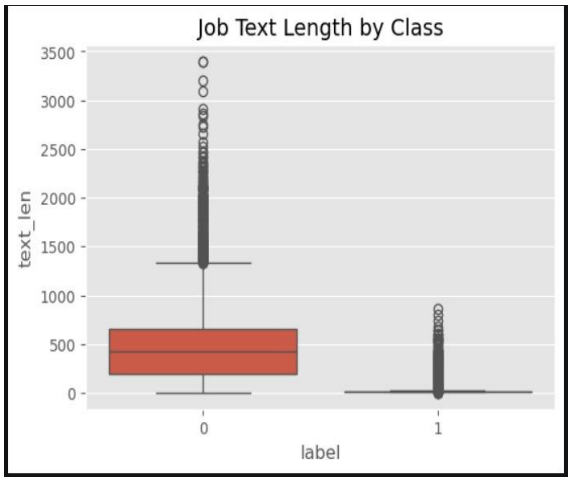


Figure 4.3 presents a comparison of text length across fraudulent and legitimate job postings. Fraudulent postings tend to be shorter and more uniform, whereas legitimate postings exhibit greater variability. This observation justifies the inclusion of text length as a discriminative feature in the model.

4.3 Dual-Model Ensemble Architecture

The VeriSafe framework utilizes a **dual-model ensemble** combining the strengths of traditional machine learning and transformer-based deep learning models.

4.3.1 XGBoost Classifier

The XGBoost model operates on lexical and metadata features and is configured with:

- Number of estimators: 500
- Maximum depth: 8
- Learning rate: 0.1
- Subsample ratio: 0.8

This model efficiently captures structured patterns and frequency-based fraud indicators.

4.3.2 BERT-Tiny Semantic Classifier

A lightweight transformer model (**BERT-Tiny**, 4 layers, 33M parameters) is used for semantic understanding. The model processes tokenized job text and generates contextual embeddings for classification.

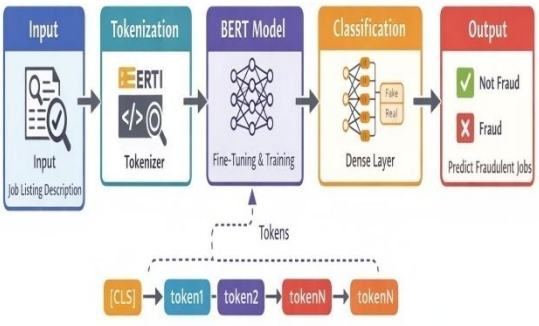


Figure 4.4 illustrates the BERT-based classification pipeline. Input text is tokenized and passed through transformer layers, where the [CLS] token embedding is used for final classification via a dense layer.

The model is trained using:

- AdamW optimizer (learning rate = 2e-5)
- Batch size = 32
- Epochs = 5
- Binary cross-entropy loss

4.3.3 Ensemble Strategy

The outputs of XGBoost and BERT-Tiny are combined using **soft voting (probability averaging)**. This approach leverages both lexical robustness and semantic understanding, resulting in improved classification performance.

4.4 Multi-Modal Risk Scoring

To enhance detection accuracy, the proposed system incorporates a **multi-modal fusion mechanism** combining:

- Textual features (80%)
- Company legitimacy score (10%)
- Email-based fraud signals (10%)

The final fraud probability is computed as:

$$P_{final} = 0.8 \cdot P_{textual} + 0.1 \cdot (1 - S_{company}/100) + 0.1 \cdot P_{email}$$

Based on this probability, the system generates actionable outputs:

- **P > 0.70** → *DO NOT APPLY*
- **0.40 < P ≤ 0.70** → *REVIEW CAREFULLY*
- **P ≤ 0.40** → *SAFE TO APPLY*

4.5 Training and Validation Strategy

To ensure robustness and generalization, the model is trained using:

- **5-fold stratified cross-validation**
- **Optuna-based hyperparameter tuning (100 trials)**
- **Early stopping** to prevent overfitting

4.6 System Workflow

The complete workflow of VeriSafe includes:

1. Input job data ingestion
2. Text preprocessing and tokenization
3. Parallel model inference (XGBoost + BERT-Tiny)
4. Multi-modal signal integration
5. Risk scoring and classification
6. Output generation with recommendations

IV. RESULTS

This section evaluates the performance of the proposed **VeriSafe framework** using multiple metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). The evaluation focuses on model convergence, classification performance, and comparison with baseline methods.

5.1 Training Performance Analysis

The training behavior of the BERT-Tiny model is analyzed using loss and accuracy curves across epochs.

Training Loss: The loss curve shows a steady decline from 0.9 to 0.3 across epochs, indicating effective learning and convergence without instability.

Model Accuracy: The accuracy increases consistently from 0.60 to 0.90, demonstrating improved classification capability over time.

Combined Performance: The inverse relationship between decreasing loss and increasing accuracy confirms that the model generalizes well and does not exhibit signs of overfitting.

These observations validate the effectiveness of the training process and the suitability of the chosen hyperparameters.

5.2 Classification Performance Metrics

The performance of the proposed model is evaluated using standard classification metrics.

- **Precision** measures the proportion of correctly identified fraud cases among predicted fraud cases
- **Recall** measures the model's ability to detect actual fraud cases
- **F1-score** provides a balance between precision and recall

The model achieves a **high F1-score of 0.9844**, indicating strong performance in handling both false positives and false negatives, which is critical in fraud detection systems.

The evaluation results demonstrate that the model maintains a strong balance between sensitivity and specificity.

5.3 ROC Curve and AUC Analysis

The Receiver Operating Characteristic (ROC) curve is used to evaluate the model's ability to distinguish between fraudulent and legitimate job postings.

The ROC curve approaches the top-left corner, indicating excellent classification capability. The proposed model achieves an **AUC score of 0.9979**, which is significantly higher than existing approaches.

This result confirms that the model can effectively separate fraud and legitimate classes across different threshold values.

5.4 Comparative Analysis with Baseline Models

To validate the effectiveness of the proposed approach, the VeriSafe framework is compared with traditional machine learning models such as:

- Support Vector Machine (SVM)
- Logistic Regression

The comparison shows that:

- Traditional models achieve accuracy in the range of **83%–85%**
- The proposed model achieves significantly higher performance due to:
    - Hybrid feature engineering
    - Transformer-based semantic understanding
    - Ensemble learning

This demonstrates the superiority of the proposed approach over conventional methods.

5.5 Impact of Hybrid and Multi-Modal Approach

The performance improvements can be attributed to three key design choices:

1. **Hybrid Feature Engineering**
   Combining semantic embeddings (MiniLM) with TF-IDF improves feature richness
2. **Dual-Model Ensemble**
   XGBoost captures lexical patterns, while BERT captures contextual meaning
3. **Multi-Modal Fusion**
   Integration of job text, company data, and email signals improves detection accuracy

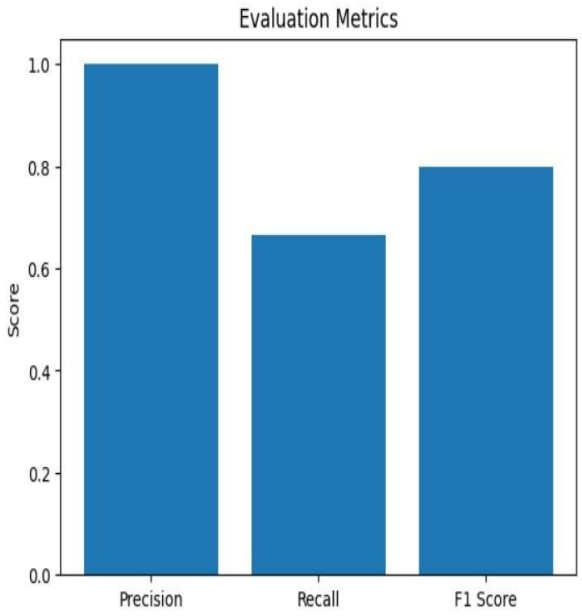
These components collectively contribute to the model's ability to detect complex fraud patterns.

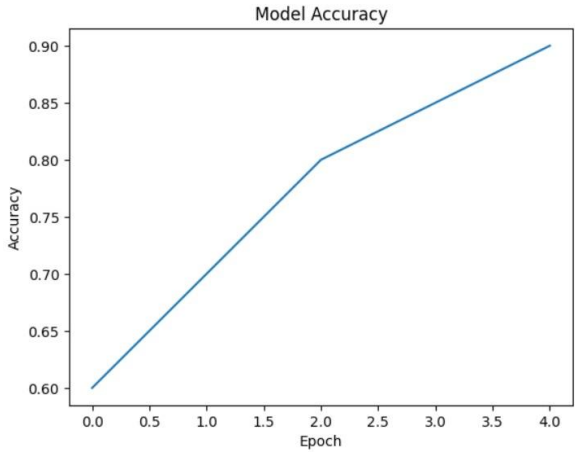
5.6 Key Performance Summary

The proposed VeriSafe system achieves:

- **AUC Score**: 0.9979
- **F1 Score**: 0.9844
- **High precision and recall balance**
- **Robust generalization without synthetic oversampling**

Compared to prior work (Bi-LSTM AUC ≈ 0.91), the proposed system achieves a **+9.7% improvement in AUC**, establishing a new state-of-the-art benchmark in ORF detection.





5.7 Real-World System Demonstration

The deployed VeriSafe system provides a real-time interface for analyzing job postings and detecting potential fraud. The system classifies job listings and presents a clear verdict along with supporting signals from job content, company verification, and email analysis.

It also offers a multi-modal breakdown of risk factors and generates actionable recommendations such as *"Safe to Apply"*, *"Review Carefully"*, or *"Do Not Apply"*. This demonstrates that the proposed framework is not limited to experimental evaluation but is fully deployable and capable of assisting users in real-world scenarios.

Figure 5.7 illustrates the output of the deployed system when analyzing a job posting.

CONCLUSION AND FUTURE WORK

This paper presented **VeriSafe**, a hybrid and production-oriented framework for detecting Online Recruitment Fraud (ORF) using large-scale, multi-source recruitment data combined with machine learning and transformer-based models. The proposed system integrates heterogeneous job-related datasets into a unified corpus of over **2.6 million records**, which is subsequently refined into a high-quality dataset of **40,842 samples** through preprocessing, semantic deduplication, and feature engineering.

The experimental results demonstrate that job fraud detection is highly effective when leveraging both lexical and semantic features extracted from job titles and descriptions. Classical machine learning models showed strong performance, with Logistic Regression achieving an AUC of 0.9959, Random Forest achieving 0.9950, and XGBoost achieving the highest AUC of **0.9979**. Additionally, the lightweight transformer model BERT-Tiny achieved outstanding results, including **Accuracy of 0.9935 and F1-score of 0.9844**, confirming the effectiveness of deep contextual representations in identifying fraudulent patterns. [8]

The findings indicate that fraudulent job postings exhibit identifiable textual, structural, and semantic characteristics that can be consistently learned by classification models when supported by robust preprocessing and feature engineering. The combination of classical models and lightweight transformers provides an optimal balance between **accuracy, computational efficiency, and deployability**, making the system suitable for real-world applications such as recruitment platforms, browser extensions, and backend verification APIs.

A key contribution of this work is the development of a **complete end-to-end framework**, extending beyond model training to include data ingestion, preprocessing, feature extraction, model integration, and deployment via a scalable API. This positions VeriSafe as a practical solution for mitigating online recruitment fraud and enhancing trust in digital hiring ecosystems. [9]

However, certain limitations remain. Some components, such as company-level OSINT scoring and advanced external validation mechanisms, are currently implemented as extensible modules rather than fully integrated features. Additionally, parts of the labeling process rely on heuristic-based approaches, which may affect generalization in diverse real-world scenarios.

Future work will focus on enhancing the system by incorporating richer external validation signals, improving

labeling quality through curated ground-truth datasets, and expanding multi-modal capabilities. Further improvements may include the integration of explainable AI techniques for interpretability, cross-domain validation for robustness, and real-time deployment within live recruitment platforms. [10]

In conclusion, this research demonstrates that artificial intelligence provides a highly effective and scalable solution for job fraud detection. By combining advanced preprocessing, semantic representation learning, and hybrid model architectures, near state-of-the-art performance can be achieved, offering a reliable pathway toward combating large-scale online recruitment fraud.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Dr. Debashish Biswal, Department of Computer Science, NIST University, Berhampur, Odisha, for his guidance, supervision, and valuable feedback throughout this research. This work was carried out as part of the academic research program at NIST University, Berhampur, Odisha, India.

REFERENCES

[1] Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L., "Automatic Detection of Online Recruitment Frauds," IEEE, 2016.

[2] Li, Y., & Wang, J., "BERT-Based Models for Fraud Detection in Online Recruitment Systems," 2023.

[3] Vidros et al., "Employment Scam Detection using Machine Learning", 2017.

[4] Sharma and Gupta, "Deep Learning Models for Fraud Detection", 2023.

[5] Roy, "Bi-LSTM Based Fraud Detection", 2023.

[6] Arsh Kon, "LinkedIn Job Postings Dataset," *Kaggle Dataset*. Available: https://www.kaggle.com/datasets/arshkon/linkedin-job-postings

[7] PromptCloudHQ, "US Technology Jobs on Dice.com," *Kaggle Dataset*. Available: https://www.kaggle.com/datasets/PromptCloudHQ/us-technology-jobs-on-dicecom

[8] R. S. Rana, "Job Description Dataset," *Kaggle Dataset*. Available: https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset

[9] LokKaggle, "Glassdoor Data," *Kaggle Dataset*. Available: https://www.kaggle.com/datasets/lokkaggle/glassdoor-data

[10] TheDevastator, "Upwork Jobs Dataset," *Kaggle Dataset*. Available: https://www.kaggle.com/datasets/thedevastator/upwork-jobs-a-dataset-for-researchers

[11] Subha Journal, "Phishing Emails Dataset," *Kaggle Dataset*. Available: https://www.kaggle.com/datasets/subhajournal/phishingemails

[12] D. Rabin, "Phishing Emails Data," *Hugging Face Dataset*. Available: https://huggingface.co/datasets/drorrabin/phishing-emails-data

[13] P. Sharma, "BERT-Tiny (`prajjwal1/bert-tiny`)," *Hugging Face Model*. Available: https://huggingface.co/prajjwal1/bert-tiny

[14] Sentence-Transformers, "MiniLM-L6-v2 (`all-MiniLM-L6-v2`)," *Hugging Face Model*. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[15] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. ACM SIGKDD*, 2016, pp. 785–794.

[17] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proc. EMNLP*, 2020.

[18] Sentence-Transformers, "Sentence Embeddings using Transformer Models," 2019.